Como lidiar con el desbalanceo de datos

joao-miranda4
23 de Marzo

Si hablamos de problemas de modelado supervisado centrados en la clasificación, podemos encontrarnos con bases de datos en las que la variable objetivo contiene clases muy desequilibradas, es decir, categorías con frecuencias muy diferentes.

Al entrenar un modelo de clasificación con la variable no balanceada, encontraremos algunos problemas. Esto sucede porque el patrón de datos de la clase dominante superará a los de la clase con menos frecuencia. Generalmente, en bases de datos que tienen una

variable objetivo desbalanceada, la clase con la frecuencia más baja es precisamente la que nos interesa predecir, lo que hace que los problemas sean aún mayores.

Como una de las clases tiene una frecuencia muy alta, el modelo construido con datos desequilibrados puede presentar una precisión muy alta y aun así no predecir correctamente ninguna observación para la clase con una frecuencia más baja. Esto puede dar la falsa impresión de que el modelo funciona bien cuando en realidad no es así.

Para solucionar estos problemas generados por una base de datos desequilibrada, podemos recurrir a dos soluciones que consisten en equilibrar los datos de la variable objetivo: ***undersampling y oversampling***.

undersampling

Es una técnica que consiste en mantener todos los datos de la clase de menor frecuencia y reducir la cantidad de los de la clase de mayor frecuencia, haciendo que las observaciones del conjunto tengan datos con la variable objetivo balanceada.

Puede ser una ventaja usar undersampling para reducir el almacenamiento de datos y el tiempo de ejecución del código, ya que la cantidad de datos será mucho menor. Una de las técnicas más utilizadas es **Near Miss**, que disminuye aleatoriamente el número de valores de la clase mayoritaria.

Algo muy interesante de Near Miss es que utiliza la menor distancia promedio de los K-vecinos más cercanos, es decir, selecciona los valores en base al método KNN (K-nearest neighbors) para reducir la pérdida de información.

Si desea saber más sobre cómo funciona la técnica Near Miss, puede consultar el artículo Enfoque de KNN para distribuciones de datos desequilibradas: un caso de estudio que implica la extracción de información.

oversampling

Es una técnica que consiste en aumentar el número de registros de la clase con menor frecuencia hasta que la base de datos tenga un número equilibrado entre las clases de la variable objetivo. Para aumentar la cantidad de registros, podemos duplicar aleatoriamente los registros de la clase con menos frecuencia. Sin embargo, esto hará que mucha información sea idéntica, lo que puede afectar el modelo.

Una ventaja de esta técnica es que no se pierde ninguna información de los registros que tenían la clase con mayor frecuencia. Esto hace que el conjunto de datos tenga muchos registros para alimentar los algoritmos de aprendizaje automático. A su vez, el tiempo de almacenamiento y procesamiento crece significativamente y existe la posibilidad de sobreajustar los datos que se han duplicado. Este sobreajuste ocurre cuando el modelo se vuelve muy bueno para predecir los resultados de los datos de entrenamiento, pero no generaliza bien los datos nuevos.

Para evitar tener demasiados datos idénticos, se puede utilizar la técnica SMOTE, que consiste en sintetizar nueva información a partir de información existente. Estos datos "sintéticos" están relativamente cerca de los datos reales, pero no son idénticos. Para obtener más información sobre cómo funciona la técnica SMOTE, puede leer el artículo SMOTE: Synthetic Minority Over-sampling Technique.

¿Cómo aplicarlos?

Ambas técnicas de balanceo se pueden aplicar utilizando la biblioteca **imbalanced-learn** que se basa en sklearn y proporciona herramientas para tratar con datos desbalanceados.

En la documentación, puedes encontrar varios ejemplos de cómo aplicar el ***undersampling y oversampling*** incluso fuera de los ejemplos presentados anteriormente. Vale la pena recordar que ambos tienen ventajas y desventajas y la aplicación de cada uno de ellos dependerá de las particularidades del problema.

João Vitor de Miranda

Licenciado en Matemáticas y posgrado en Data Science and Analytics. Con conocimientos en Matemáticas, Estadística, Excel, Python, R y SQL/NoSQL.

Paga en moneda local en los siguientes países

ANUAL

US$79,90

un solo pago de US$79,90

- 218 cursos
- Videos y actividades 100% en Español
- Certificado de participación
- Estudia las 24 horas, los 7 días de la semana
- Foro y comunidad exclusiva para resolver tus dudas
- Acceso a todo el contenido de la plataforma por 12 meses

¡QUIERO EMPEZAR A ESTUDIAR!

BLOG

PROGRAMACIÓN
FRONT END
DATA SCIENCE
INNOVACIÓN Y GESTIÓN
DEVOPS

AOVS Sistemas de Informática S.A
CNPJ 05.555.382/0001-33

SÍGUENOS EN NUESTRAS REDES SOCIALES

ALIADOS

Empresa participante do SCALE UP ENDEAVOR

En Alura somos unas de las Scale-Ups seleccionadas por Endeavor, programa de aceleración de las empresas que más crecen en el país.

Google for Startups Alumni 2021

Fuimos unas de las 7 startups seleccionadas por Google For Startups en participar del programa Growth Academy en 2021

POWERED BY

CURSOS

Cursos de Programación

Lógica de Programación | Java

Cursos de Front End

HTML y CSS | JavaScript | React

Cursos de Data Science

Data Science | Machine Learning | Excel | Base de Datos | Data Visualization | Estadística

Cursos de DevOps

Docker | Linux

Cursos de Innovación y Gestión

Productividad y Calidad de Vida | Transformación Ágil | Marketing Analytics | Liderazgo y Gestión de Equipos | Startups y Emprendimiento